Audio-Visual-Tactile integration in Speech Perception

Donald Derrick, Doreen Hansmann, Zoe Haws & Catherine Theys
University of Canterbury

Behavioural audio-visual research has shown enhancement[1] and interference[2] in speech perception, as has behavioural audio-tactile research[3]. However, to date, we have not encountered any experimental behavioural research into tri-modal integration in speech perception. (But see Alcorn[4] for clinical techniques incorporating both vision and touch.) Based on the relative influence of visual and aero-tactile stimuli, we expect cumulative effects of both, with the most influence from auditory information, then visual information[1], and lastly airflow[3]. Here we present a two-way forced-choice study of tri-modal integration in speech perception, showing the effects of both congruent and incongruent stimuli on accurate identification of auditory speech-in-noise.

Hypotheses: 1) Participants will more accurately identify audio stimuli in noisy environments if they are paired with matching multi-sensory stimuli, with the multisensory enhancements stacking; tactile less enhancing than visual stimuli, which is less enhancing than the combination of both visual and tactile stimuli. 2) Mismatching multisensory stimuli will interfere with accurate identification of audio stimuli, with the tactile mismatch least interfering, followed by visual, followed by the most interfering combination of mismatched visual and tactile stimuli.

Methods: Audiovisual recordings of one female speaker producing "pa" [p^{h}a] and "ga" [ka] were labeled segmented in PRAAT[5]. Four of each were selected and used as the base stimuli based on duration, f0, and the elimination of stimuli with eye-blinks. All of the speech tokens were randomly superimposed over a 10-second loop of sound using an automated process in R[6]. This technique insures that the audio tokens are masked by noise that matches the long-term spectrum of the underlying speech[7,8]. The audio and noise were superimposed, giving a range of -20 to +10 decibel (dB) signal-to-noise ratios (SNR) at 0.1 dB increments.

Sixteen New Zealand English speakers participated; the final experiment will have 30-40 participants based on our power-analysis. Participants were seated in a sound-attenuated booth with a screen 1 meter in front of them, an air-puff system[9] directed at their suprasternal notch, and headphones placed over their ears. They were asked to identify between [p^{h}a] and [ka], beginning with a +1 dB SNR, following a QUEST[10] staircase along 32 choices for each condition. Conditions consisted of matched and mismatched audio-visual and audio-visual-tactile stimuli for a total of 12 conditions (see Figure 1).

Results: The results, as seen in Figure 1, show a trend that congruous video and tactile stimuli enhance, and incongruous video and tactile stimuli interfere with, identification of auditory [p^{h}a] and [ka], as predicted in Hypothesis 1. However, GLMM tests show only the visual enhancement and interference are statistically significant (t-value = 8.060); in contrast the puff enhancement and interference are not significant (t-value = 1.215).

Discussion: The current number of participants provides data that is too under-powered to allow for effective elimination of type II statistical errors. However, the current results already show a data trend towards tri-modal integration in speech perception. The results also show that audio-visual integration has about a seven times larger effect range (4.65 dB SNR) than audio-tactile integration (0.66 dB SNR), so continued research into tri-modal integration in speech perception will require either greater air-flow or more repetitions to avoid under-powered analyses.

References (continued on page 2):

[1] Sumby, W. H., & Pollack, I. 1953. Visual Contribution to Speech Intelligibility in Noise, *Journal of the Acoustical Society of America*, 26, 212-215.

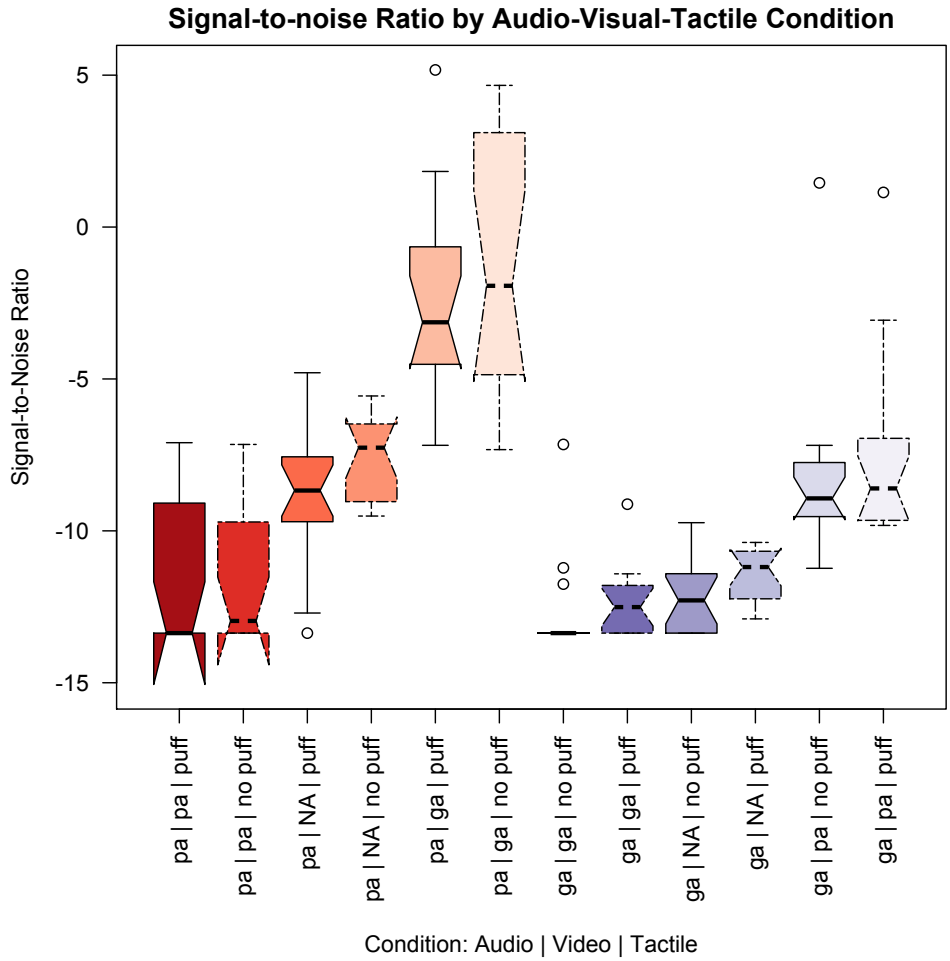

Figure 1. *Notched boxplots of Signal-to-Noise Ratio by Audio-Visual-Tactile Condition (NA = no video)*

[2] McGurk, H., & MacDonald, J., 1976. Hearing lips and seeing voices, *Nature*, 264:746-748.
[3] Gick, B., & Derrick, D. 2009. Aero-tactile integration in speech perception. *Nature*, 462, 502-504.
[4] Alcorn, S. 1932. The Tadoma method, *Volta Review,* 34, 195-198.
[5] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International,* **5:9/10**, 341-345.
[6] R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[7] Jansen, S., Luts, H., Wagener, K. C., Frachet, B., & Wouters, J. 2010. The French digit triplet test: A hearing screening tool for speech intelligibility in noise. *International Journal of Audiology*, 49(5), 378–387.
[8] Smits C., Kapteyn,T.S., & Houtgast, T. 2004. Development and validation of an automatic speech-in-noise screening test by telephone, *International Journal of Audiology*, 43(1), 15– 28.
[9] Derrick, D., De Rybel, T., O'Beirne, G. A., Hay, J. 2014. Listen with your skin: Aerotak speech perception enhancement system, in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, 1484-1485
[10] Watson, A. B. 1983. QUEST: A Bayesian adaptive psychometric method, *Perception & Psychophysics*, 33(2), 113-120